

日 本 国 特 許 庁  
JAPAN PATENT OFFICE

30.07.03 #2

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日  
Date of Application:

2002年 7月30日

出 願 番 号  
Application Number:

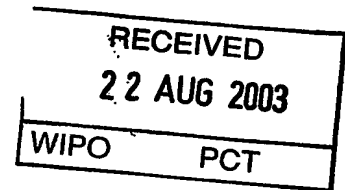
特願2002-221698

[ ST.10/C ]:

[ JP2002-221698 ]

出 願 人  
Applicant(s):

ソニー株式会社

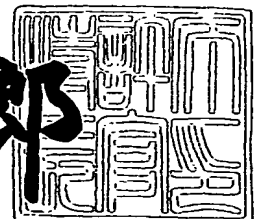


**PRIORITY  
DOCUMENT**  
SUBMITTED OR TRANSMITTED IN  
COMPLIANCE WITH RULE 17.1 (a) OR (b)

2003年 6月 9日

特 許 庁 長 官  
Commissioner,  
Japan Patent Office

太田 信一郎



出証番号 出証特2003-3044520

【書類名】 特許願

【整理番号】 0290377102

【提出日】 平成14年 7月30日

【あて先】 特許庁長官殿

【国際特許分類】 H04N 5/44

【発明者】

    【住所又は居所】 東京都品川区北品川 6 丁目 7 番 3 5 号 ソニー株式会社  
内

    【氏名】 木村 仁史

【発明者】

    【住所又は居所】 東京都品川区北品川 6 丁目 7 番 3 5 号 ソニー株式会社  
内

    【氏名】 大沼 顕介

【発明者】

    【住所又は居所】 東京都品川区北品川 6 丁目 7 番 3 5 号 ソニー株式会社  
内

    【氏名】 市岡 秀俊

【特許出願人】

    【識別番号】 000002185

    【氏名又は名称】 ソニー株式会社

【代理人】

    【識別番号】 100122884

    【弁理士】

    【氏名又は名称】 角田 芳末

    【電話番号】 03-3343-5821

【選任した代理人】

    【識別番号】 100113516

    【弁理士】

    【氏名又は名称】 磯山 弘信

【電話番号】 03-3343-5821

【手数料の表示】

【予納台帳番号】 176420

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 0206460

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 キーワードの自動抽出装置及び方法

【特許請求の範囲】

【請求項 1】 コンテンツのタイトル文字列情報から、サブジャンルを示す文字列が登録されている第 1 のキーワード辞書を用いてキーワードの抽出を行う第 1 の抽出手段と、

前記コンテンツの詳細文字列情報から、人名が登録されている第 2 のキーワード辞書を用いたキーワードの抽出と、字種切り法を利用したキーワードの抽出とを行う第 2 の抽出手段と

を備えたことを特徴とするキーワード自動抽出装置。

【請求項 2】 請求項 1 に記載のキーワード自動抽出装置において、

前記第 1 の抽出手段は、第 1 のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、所定の除外文字列辞書に登録されている文字列を除外した部分からキーワードを抽出することを特徴とするキーワード自動抽出装置。

【請求項 3】 請求項 1 に記載のキーワード自動抽出装置において、

前記第 1 の抽出手段は、第 1 のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、平仮名、カタカナ、漢字、数字、アルファベット以外の特殊文字で区切られている文字列をキーワードとして抽出することを特徴とするキーワード自動抽出装置。

【請求項 4】 請求項 1 に記載のキーワード自動抽出装置において、

前記第 2 の抽出手段は、前記第 2 のキーワード辞書を用いてキーワードを抽出した前記詳細文字列情報の残りの部分のうち、所定の除外文字列辞書に登録されている文字列を除外した部分から、字種切り法を利用したキーワードの抽出を行うことを特徴とするキーワード自動抽出装置。

【請求項 5】 請求項 1 に記載のキーワード自動抽出装置において、

前記第 2 の抽出手段は、字種切り法を利用しつつ、カタカナとアルファベットとを同一字種として扱うとともに、‘・’（中点）は、その直前の文字がカタカナ、アルファベットである場合にはそれぞれカタカナ、アルファベットとして扱うことを特徴とするキーワード自動抽出装置。

【請求項6】 請求項1に記載のキーワード自動抽出装置において、

前記第2のキーワード辞書をネットワーク経由でダウンロードする手段をさらに備え、前記第2の抽出手段は前記ダウンロードされた第2のキーワード辞書を用いることを特徴とするキーワード自動抽出装置。

【請求項7】 コンテンツのタイトル文字列情報から、サブジャンルを示す文字列が登録されている第1のキーワード辞書を用いてキーワードの抽出を行う第1のステップと、

前記コンテンツの詳細文字列情報から、人名が登録されている第2のキーワード辞書を用いたキーワードの抽出と、字種切り法を利用したキーワードの抽出とを行う第2のステップと

を有することを特徴とするキーワード自動抽出方法。

【請求項8】 請求項7に記載のキーワード自動抽出方法において、

前記第1のステップで、第1のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、所定の除外文字列辞書に登録されている文字列を除外した部分からキーワードを抽出することを特徴とするキーワード自動抽出方法。

【請求項9】 請求項7に記載のキーワード自動抽出方法において、

前記第1のステップで、第1のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、平仮名、カタカナ、漢字、数字、アルファベット以外の特殊文字で区切られている文字列をキーワードとして抽出することを特徴とするキーワード自動抽出方法。

【請求項10】 請求項7に記載のキーワード自動抽出方法において、

前記第2のステップで、前記第2のキーワード辞書を用いてキーワードを抽出した前記詳細文字列情報の残りの部分のうち、所定の除外文字列辞書に登録されている文字列を除外した部分から、字種切り法を利用したキーワードの抽出を行うことを特徴とするキーワード自動抽出方法。

【請求項11】 請求項7に記載のキーワード自動抽出方法において、

前記第2のステップで、字種切り法を利用しつつ、カタカナとアルファベットとを同一字種として扱うとともに、‘・’（中点）は、その直前の文字がカタカナ、アルファベットである場合にはそれぞれカタカナ、アルファベットとして扱

うことを特徴とするキーワード自動抽出方法。

【請求項 1 2】 請求項 7 に記載のキーワード自動抽出方法において、

前記第 2 のキーワード辞書をネットワーク経由でダウンロードするステップをさらに有し、前記第 2 のステップでは前記ダウンロードした第 2 のキーワード辞書を用いることを特徴とするキーワード自動抽出方法。

【発明の詳細な説明】

【0 0 0 1】

【発明の属する技術分野】

本発明は、E P G (Electronic Program Guide: 電子番組ガイド) 情報のようなコンテンツのタイトル文字列情報及び詳細文字列情報から、キーワードを自動的に抽出する装置及び方法に関する。

【0 0 0 2】

【従来の技術】

近年本格化しているデジタルテレビジョン放送では、番組の映像・音声データとともに、番組のタイトルを示す情報（タイトル文字列情報）や番組の詳細を説明する情報（詳細文字列情報）や番組のジャンルを示す情報等を含んだ E P G 情報が、放送局から送信される。デジタル放送に対応したテレビジョン受信機では、この E P G 情報に基づいて画面上に電子番組ガイドを表示させることができる。

【0 0 0 3】

また、アナログテレビジョン放送でも、こうした E P G 情報が送信されているものがある。

【0 0 0 4】

ユーザーは、見たい番組を検索する場合、この電子番組ガイドを利用して、大まかなジャンル（例えばスポーツ、ドラマ等）を選んだ後、タイトルから検索したり、詳細文字列情報を読んで検索したりしている。

【0 0 0 5】

しかし、番組のタイトルの付け方は千差万別であるので、タイトルからの検索はユーザーにとって必ずしも行いやすいものではない。また、番組の詳細文字列

情報は文章の形式で記述されており何ページにも亘ることが少なくないので、詳細文字列情報からの検索もユーザーにとって面倒である。

【0006】

これに対し、例えば芸能人名等のキーワードを用いて番組を検索できるようにすれば、ユーザーにとって検索が非常に容易になる。しかるに、現在放送局から送信されるEPG情報には、キーワードは独立して含まれていない。そのため、キーワードを用いた検索を可能にするためには、EPG情報からキーワードを抽出することが必要になる。

【0007】

従来、このキーワードの抽出方法としては、テレビジョン受信機に表示された電子番組ガイド中の詳細文字列情報の文章の中から、ユーザーが、キーワードとして決定したい文字列の先頭及び末尾の語をカーソル等で指定するという方法が存在していた。

【0008】

【発明が解決しようとする課題】

しかし、この従来の抽出方法では、ユーザーが自らキーワード指定のための操作を行わなければならないので、煩雑であるとともに、多数のキーワードを短時間に抽出することは困難である。

【0009】

他方、一般的なキーワードの自動抽出方法としては、日本語形態要素解析という方法が知られている。しかし、この方法は、プログラムのサイズや使用する辞書のサイズが非常に大きいとともに、CPUに大きな負荷をかけてしまう。したがって、テレビジョン受信機のようにCPUの処理能力やメモリの容量がさほど大きくない家電製品でこの方法を用いることは、極めて非効率的である。

【0010】

さらに、一般的なキーワードの自動抽出方法としては、字種切り法という方法も知られている。この方法は、漢字・カタカナ・平仮名・アルファベット・数字等の字種の違いを検出することによってキーワードを抽出するものである。しかし、この字種切り法だけでは、番組を検索するためのキーワードの抽出を精度よ

く行うことはできない。すなわち、苗字が漢字で名前が平仮名やカタカナの芸能人名（例えば‘石田あかり’というような名称）は、苗字と名前とが分割されてしまうので抽出することができない。また、名前がアルファベットで表記され苗字がカタカナで表記された外国人名や名前と苗字との間に‘・’（中点）が挿入された外国人名（例えば‘B・ドゥーリー’というような名称）も、名前と苗字とが分割されてしまうので抽出することができない。

#### 【0011】

本発明は、上述の点に鑑み、CPUの処理能力やメモリの容量がさほど大きくない家電製品でも、EPG情報のようなコンテンツのタイトル文字列情報及び詳細文字列情報から、ユーザーがコンテンツを検索するためのキーワードを、効率よく且つ精度よく自動的に抽出できるようにすることを課題としてなされたものである。

#### 【0012】

##### 【課題を解決するための手段】

この課題を解決するために、本出願人は、コンテンツのタイトル文字列情報から、サブジャンルを示す文字列が登録されている第1のキーワード辞書を用いてキーワードの抽出を行う第1の抽出手段と、このコンテンツの詳細文字列情報から、人名が登録されている第2のキーワード辞書を用いたキーワードの抽出と、字種切り法を利用したキーワードの抽出とを行う第2の抽出手段とを備えたキーワード自動抽出装置を提案する。

#### 【0013】

このキーワード自動抽出装置では、コンテンツのタイトル文字列情報（例えばデジタルテレビジョン放送ではEPG情報中のタイトル文字列情報）からは、サブジャンルを示す文字列が登録されている第1のキーワード辞書を用いてキーワードが抽出される。

#### 【0014】

また、コンテンツの詳細文字列情報（例えばデジタルテレビジョン放送ではEPG情報中の詳細文字列情報）からは、人名が登録されている第2のキーワード辞書を用いてキーワードが抽出されるとともに、字種切り法を利用したキーワー



ドの抽出も行われる。その際、第2のキーワード辞書に登録されている人名であれば、苗字が漢字で名前が平仮名やカタカナの人名もキーワードとして抽出される。また、第2のキーワード辞書に登録されていない人名でも、字種切り法を利用することによってキーワードとして抽出される。

#### 【0015】

このように、タイトル文字列情報からのキーワードの抽出と詳細文字列情報からのキーワードの抽出とを、それぞれの情報に合せて互いに異なるキーワード辞書とルール（字種切り法を利用するか否か等）で行うことにより、小さなサイズのプログラムや辞書で精度よくキーワードを抽出することができる。

#### 【0016】

これにより、CPUの処理能力やメモリの容量がさほど大きくない家電製品でも、EPG情報のようなコンテンツのタイトル文字列情報及び詳細文字列情報から、ユーザーがコンテンツを検索するためのキーワードを、効率よく且つ精度よく自動的に抽出することができるようになる。

#### 【0017】

なお、このキーワード自動抽出装置において、一例として、第1の抽出手段は、第1のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、所定の除外文字列辞書に登録されている文字列を除外した部分からキーワードを抽出することが好適である。

#### 【0018】

それにより、タイトルに含まれていることのある文字列のうち、コンテンツを検索するためには不適切な（一般的過ぎるような）文字列がキーワードに含まれることを防止することができる。したがって、ユーザーは、抽出されたキーワードを用いて、コンテンツを一層効率的に検索することができるようになる。

#### 【0019】

さらに、このキーワード自動抽出装置において、一例として、第1の抽出手段は、第1のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、平仮名、カタカナ、漢字、数字、アルファベット以外の特殊文字で区切られている文字列をキーワードとして抽出することが好適である。

## 【0020】

それにより、こうした特殊文字で区切られていないタイトルについては、そのタイトルに含まれる複数の文字列がばらばらのキーワードとして抽出されることなく、そのタイトルそのものがそのままの形でキーワードとして抽出されるようになる。

## 【0021】

こうした特殊文字で区切られていないタイトルは、そのタイトルに含まれる個々の文字列は意味が広すぎてコンテンツ検索のためのキーワードとしてあまり役立たず（検索結果が非常に多くなり）、タイトルそのものとしてはじめてコンテンツの効率的な検索のためのキーワードとして役立つことが多い。したがって、ユーザーは、抽出されたキーワード（タイトルそのもの）を用いて、コンテンツを一層効率的に検索することができるようになる。

## 【0022】

また一方では、特殊文字で区切られているタイトルについては、特殊文字で区切られている個々の文字列がそれぞれキーワードとして抽出されるようになる。

## 【0023】

特殊文字（例えばスペースや‘×’等）で区切られているタイトルは、その特殊文字で区切られている個々の文字列がそれぞれコンテンツ検索のためのキーワードとして役立ち、タイトルそのものとしては限定されすぎてコンテンツ検索のためのキーワードとしてあまり役立たない（検索結果がゼロまたは非常に少なくなる）ことが多い。したがって、ユーザーは、抽出されたキーワード（特殊文字で区切られている個々の文字列）を用いて、やはりコンテンツを一層効率的に検索することができるようになる。

## 【0024】

さらに、このキーワード自動抽出装置において、一例として、第2の抽出手段は、第2のキーワード辞書を用いてキーワードを抽出した詳細文字列情報の残りの部分のうち、所定の除外文字列辞書に登録されている文字列を除外した部分から、字種切り法を利用したキーワードの抽出を行うことが好適である。

## 【0025】

それにより、詳細文字列情報に含まれていることのある文字列のうち、コンテンツを検索するためには不適切な文字列がキーワードに含まれることを防止することができる。したがって、ユーザーは、抽出されたキーワードを用いて、コンテンツを一層効率的に検索することができるようになる。

## 【0026】

さらに、このキーワード自動抽出装置において、一例として、第2の抽出手段は、字種切り法を利用しつつ、カタカナとアルファベットとを同一字種として扱うとともに、‘・’（中点）は、その直前の文字がカタカナ、アルファベットである場合にはそれぞれカタカナ、アルファベットとして扱うことが好適である。

## 【0027】

それにより、名前がアルファベットで表記され苗字がカタカナで表記された外国人名や名前と苗字との間に‘・’（中点）が挿入された外国人名も、キーワードとして抽出することができるようになる。

## 【0028】

さらに、このキーワード自動抽出装置において、第2のキーワード辞書をネットワーク経由でダウンロードする手段をさらに備え、第2の抽出手段はこのダウンロードされた第2のキーワード辞書を用いることが好適である。

## 【0029】

それにより、第2のキーワード辞書として、最新の辞書（最近有名になったばかりの人の名称も登録されている辞書）を用いてキーワードを抽出することができるようになる。

## 【0030】

次に、本出願人は、コンテンツのタイトル文字列情報から、サブジャンルを示す文字列が登録されている第1のキーワード辞書を用いてキーワードの抽出を行う第1のステップと、このコンテンツの詳細文字列情報から、人名が登録されている第2のキーワード辞書を用いたキーワードの抽出と、字種切り法を利用したキーワードの抽出とを行う第2のステップとを有するキーワード自動抽出方法を提案する。

## 【0031】

このキーワード自動抽出方法によれば、前述の本発明に係るキーワード自動抽出装置について説明したのと全く同様にして、CPUの処理能力やメモリの容量がさほど大きくない家電製品でも、EPG情報のようなコンテンツのタイトル文字列情報及び詳細文字列情報から、ユーザーがコンテンツを検索するためのキーワードを、効率よく且つ精度よく自動的に抽出することができるようになる。

【0032】

【発明の実施の形態】

以下、デジタルテレビジョン放送の番組を記録・再生する装置に本発明を適用した例について、図面を用いて説明する。

【0033】

図1は、本発明を適用した番組記録再生装置を含むデジタルテレビジョン放送受信システムの概要を示す図である。テレビジョン放送局から送信されたデジタル放送信号が、アンテナ1で受信されて番組記録再生装置2に入力する。番組記録再生装置2は、ディスプレイ及びスピーカを含む表示装置3に接続されるとともに、インターネット4に接続されている。

【0034】

図2は、番組記録再生装置2のハードウェア構成を示すブロック図である。この番組記録再生装置2では、チューナ11、復調器12、デスクランブラ13、多重分離器14が順に接続されるとともに、多重分離器14に対して映像デコーダ15、映像信号処理回路17と音声デコーダ16、D/A変換器18とがそれぞれ順に接続されている。

【0035】

また、チューナ11～D/A変換器18、CPU19、ROM20、メインメモリ(RAM)21、フラッシュメモリ22、リモートコントローラ用のインターフェース23、HDD(ハードディスクドライブ)用のインターフェース24、インターネット接続用の通信インターフェース25が、互いにシステムバス26で結ばれている。インターフェース24には、テレビ番組を録画するためのHDD(ハードディスクドライブ)27が接続されている。

【0036】

この番組記録再生装置 2 に付属したリモートコントローラ（以下リモコンと呼ぶ）28 には、通常のデジタル放送用テレビジョン受信機に付属したリモコンにおけるのと同じ各種の操作釦（電源釦や、選局釦や、録画予約釦や、再生釦や、E P G 画面上で選択を行うための方向キーや決定キー等）が設けられている。

#### 【0037】

テレビ番組の視聴時には、番組記録再生装置 2 に入力したデジタル放送信号は、リモコン 28 の選局操作に基づいてチューナ 11 で周波数帯を選択された後、復調器 12 で復調され、デスクランブラ 13 でスクランブルを解かれた後、多重分離器 14 で、複数チャンネル分の番組の映像・音声データの packets や E P G 情報の packets に分離される。

#### 【0038】

この複数チャンネル分のテレビ番組の映像・音声の packets のうち、リモコン 28 の選局操作に基づいて抽出した 1 チャンネル分の packets の映像、音声データが、それぞれ映像デコーダ 15、音声デコーダ 16 により M P E G - 2 V i d e o、M P E G - 2 A u d i o で復号される。また、E P G 情報の packets は C P U 19 に送られる。

#### 【0039】

そして、映像デコーダ 15 で復号された映像信号や、E P G 情報を用いて C P U 19 で作成された電子番組ガイド表示用の映像信号が、映像信号処理回路 17 で N T S C 方式への変換やミキシング等を施され、映像出力端子 29 から出力して図 1 の表示装置 3 に送られる。

#### 【0040】

また、音声デコーダ 16 で復号された音声信号が、D / A 変換器 18 でアナログ変換され、音声出力端子 30 から出力して図 1 の表示装置 3 に送られる。

#### 【0041】

C P U 19 は、ROM 20 に格納されたプログラムやデータに基づき、メインメモリ 21 をワーキングメモリとして用いてこの番組記録再生装置 2 全体を制御する。

#### 【0042】

CPU19が行う処理には、リモコン28の選局操作に基づくテレビ番組の視聴時の処理や、リモコン28の録画予約操作に基づくHDD27へのテレビ番組の録画処理の他に、キーワードの自動抽出処理がある。

【0043】

ROM20には、このキーワードの自動抽出処理で用いるための辞書として、タイトル用キーワード辞書、タイトル用除外文字列辞書、詳細情報用キーワード辞書及び詳細情報用除外文字列辞書が格納されている。

【0044】

タイトル用キーワード辞書には、‘プロ野球’，‘ゴルフ’，‘サッカー’，‘温泉’，‘囲碁’，‘将棋’，‘映画’等のサブジャンル（EPG情報中のジャンル情報による‘スポーツ’といったような大まかなジャンルよりも細かいジャンル）を示す文字列や、‘恋’，‘愛’といった文字列や、プロ野球の球団名の文字列のように、番組のタイトルに含まれていることの多い文字列のうち、番組を検索するために有効且つ重要な文字列が登録されている。

【0045】

タイトル用除外文字列辞書には、‘映画’，‘BS’，番組表独特の記号（例えば、ニュース番組を表す、Nを四角の枠で囲んだ記号）といったように、番組のタイトルに含まれていることのある文字列のうち、番組を検索するためのキーワードとしては一般的過ぎる文字列が登録されている。

【0046】

詳細情報用キーワード辞書には、テレビジョン番組に登場することの多い有名人（芸能人，スポーツ選手，政治家，文化人等）の名称のうち、平仮名のみ，平仮名と漢字との組み合わせ，平仮名とカタカナとの組み合わせ，漢字とカタカナとの組み合わせ，2文字以下の漢字のみ，6文字以上の漢字のみの名称の文字列がそれぞれ登録されている。また、詳細情報用キーワード辞書には、例えば‘温泉’というような、EPG情報中の詳細文字列情報に含まれていることの多い文字列のうち、人名以外の文字列であって番組を検索するためのキーワードとして適切な文字列も登録されている。

【0047】

詳細情報用除外文字列辞書には、‘ゲスト’，‘以上’，‘監督’のように、E P G情報中の詳細文字列情報に含まれていることの多い文字列のうち、番組を検索するためのキーワードとしては不適切な文字列が登録されている。

【0048】

なお、CPU19は、詳細情報用キーワード辞書に関しては、インターネット経由で専用のサイトから最新のもの（最近有名になったばかりの人の名称等が登録されているもの）をダウンロードしてフラッシュメモリ22にも記憶させる。

【0049】

また、CPU19は、キーワードの自動抽出処理を行う前提として、ユーザーの選局操作時やユーザーの録画予約操作に基づく録画時に多重分離器14から送られたE P G情報のパケットをフラッシュメモリ22に記憶させる。

【0050】

図3，図4は、CPU19が実行するキーワードの自動抽出処理を示すフローチャートである。このうち、図3は、タイトル文字列情報からキーワードを抽出する処理であり、最初に、フラッシュメモリ22に記憶させたE P G情報の中から、タイトル文字列情報を取り出す（ステップS1）。

【0051】

続いて、そのタイトル文字列情報が示す複数の番組のタイトルから、タイトル用キーワード辞書に登録されている文字列（‘ゴルフ’，‘サッカー’，‘温泉’，‘囲碁’，‘将棋’，‘映画’といったサブジャンルを示す文字列等）を探す。そして、それらの番組のタイトルのうち、このタイトル用キーワード辞書に登録されている文字列が含まれているタイトルの文字列全体を、キーワードの抽出対象とする（ステップS2）。

【0052】

続いて、ステップS2でキーワードの抽出対象としたタイトルのうち、タイトル用除外文字列辞書に登録されている文字列（‘映画’，‘BS’等）の部分をスペースで置換する（ステップS3）。

【0053】

続いて、ステップS3を経たタイトルの文字列から、図5に示すようなタイト

ル用の抽出ルールでキーワードを抽出する（ステップS4）。

【0054】

このタイトル用抽出ルールでは、そのタイトルの文字列が平仮名、カタカナ、漢字、数字、アルファベット以外の特殊文字（スペースや×や「」等）で区切られていない場合は、そのタイトルの文字列をそのままキーワードとして抽出する。他方、そのタイトルの文字列がこうした特殊文字で区切られている場合は、特殊文字で区切られている各文字列のうちの2文字以上の文字列をそれぞれキーワードとして抽出する。

【0055】

但し、‘.’（中点）は特殊文字としては扱わない。そして、キーワードとして抽出した文字列の先頭か末尾に‘.’（中点）が存在する場合には、‘.’（中点）を除いた部分をキーワードとする。

【0056】

最後に、ステップS4で抽出したキーワードを、タイトル文字列情報中のキーワードのリストとしてフラッシュメモリ22に記憶させる（ステップS5）。

【0057】

次に、図4は、詳細文字列情報からキーワードを抽出する処理であり、最初に、フラッシュメモリ22に記憶させたEPG情報の中から、詳細文字列情報を取り出す（ステップS11）。

【0058】

続いて、その詳細文字列情報から、詳細情報用キーワード辞書に登録されている文字列（有名な人の名称等）を探す。そして、その詳細文字列情報のうち、この詳細情報用キーワード辞書に登録されている文字列をキーワードとして抽出するとともに、その文字列の部分を半角スペースで置換する（ステップS12）。

【0059】

続いて、ステップS12を経た詳細文字列情報の文字列のうち、詳細情報用除外文字列辞書に登録されている文字列（‘ゲスト’，‘以上’，‘監督’等）の部分半角スペースで置換する（ステップS13）。

【0060】



続いて、ステップ S 1 3 を経た詳細文字列情報の文字列から、図 6 に示すような詳細文字列情報用の抽出ルールでキーワードを抽出する（ステップ S 1 4）。

#### 【0061】

この詳細文字列情報用抽出ルールでは、基本的には、平仮名とカタカナと漢字と数字とアルファベットとその他の字種の文字とを互いに分離する字種切り法を利用する。

#### 【0062】

但し、カタカナとアルファベットとは同一の字種として扱う（分離しない）。また、‘・’（中点）は、その直前の文字がカタカナ、アルファベットである場合にはそれぞれカタカナ、アルファベットとして扱う（分離しない）。

#### 【0063】

そして、分離した各文字列のうち、平仮名のみの文字列、2文字以下の漢字のみの文字列、6文字以上の漢字のみの文字列を除いた文字列を、それぞれキーワードとして抽出する。但し、キーワードとして抽出した文字列の先頭か末尾に‘・’（中点）が存在する場合には、‘・’（中点）を除いた部分をキーワードとする。

#### 【0064】

最後に、ステップ S 1 2 で抽出したキーワードと、ステップ S 1 4 で抽出したキーワードを、詳細文字列情報中のキーワードのリストとしてフラッシュメモリ 2 2 に記憶させる（ステップ S 1 5）。

#### 【0065】

次に、この番組記録再生装置 2 において番組検索のためのキーワードが抽出される様子を、具体例を挙げて説明する。

#### 【0066】

ユーザーの選局操作時やユーザーの録画予約操作に基づく録画時に多重分離器 1 4 から送られてフラッシュメモリ 2 2 に記憶された E P G 情報中のタイトル文字列情報に、例えば次のようなタイトルが含まれていたとする（但し□□、△△はプロ野球チーム名である）。

愛のから騒ぎ

プロ野球中継 □□×△△

BS映画「スペース・ウォーズ」

【0067】

すると、図3の処理では、‘愛’，‘プロ野球’，‘映画’という文字列がタイトル用キーワード辞書に登録されているので、ステップS2で、これらの各タイトルについて、それぞれそのタイトルの文字列全体がキーワードの抽出対象となる。

【0068】

そして、これらのタイトルのうち、BS映画「スペース・ウォーズ」については、ステップS3で、‘BS’の部分と‘映画’の部分とがスペースで置換される。

【0069】

また、これらのタイトルのうち、プロ野球中継 □□×△△については、‘プロ野球中継’と‘□□’との間にスペース（特殊記号）が存在し、‘□□’と‘△△’との間にも×（特殊記号）が存在するので、ステップS4で、文字列‘プロ野球中継’，‘□□’，‘△△’がそれぞれキーワードとして抽出される。

【0070】

また、これらのタイトルのうち、‘BS’，‘映画’の部分をスペース置換された「スペース・ウォーズ」については、「」（特殊記号）で区切られており、また‘・’（中点）は特殊記号として扱わないので、ステップS4で、本来の映画のタイトルそのものである‘スペース・ウォーズ’がキーワードとして抽出される。

【0071】

また、これらのタイトルのうち、愛のから騒ぎは、特殊記号で区切られていないので、ステップS4で、タイトルそのものである‘愛のから騒ぎ’がキーワードとして抽出される。

【0072】

したがって、ステップS5では、以下の文字列が番組検索用のキーワードとしてフラッシュメモリ22に記憶される（前述のように□□，△△はプロ野球チー

ム名である)。

愛のから騒ぎ

プロ野球中継

□□

△△

スペース・ウォーズ

【0073】

このようにして、愛のから騒ぎ、スペース・ウォーズというように特殊文字で区切られていないタイトルについては、図3の処理により、そのタイトルに含まれる複数の文字列がばらばらのキーワードとして抽出されることなく、そのタイトルそのものがそのままの形でキーワードとして抽出される。

【0074】

こうした特殊文字で区切られていないタイトルは、そのタイトルに含まれる‘愛’，‘スペース’等の個々の文字列は意味が広すぎて番組検索のためのキーワードとしてあまり役立たず（検索結果が非常に多くなり）、タイトルそのものとしてはじめて番組の効率的な検索のためのキーワードとして役立つことが多い。したがって、ユーザーは、抽出されたキーワード（タイトルそのもの）を用いて、番組を効率的に検索することができるようになる。

【0075】

また、スペース・ウォーズという映画のタイトル文字列については、タイトル文字列情報においてこのタイトルに付加されていた‘BS’，‘映画’というような番組検索のためには一般的過ぎる文字列がキーワードに含まれていないとともに、このタイトル文字列情報においてこのタイトルを囲っていた「」もキーワードに含まれていない。したがって、ユーザーは番組を効率的に検索することができるようになる。

【0076】

また一方では、プロ野球中継 □□×△△というように特殊文字（スペースや‘×’）で区切られているタイトルについては、図3の処理により、特殊文字で区切られている個々の文字列であるプロ野球中継、□□、△△がそれぞれキーワ

ードとして抽出される。

【0077】

こうした特殊文字で区切られているタイトルは、その特殊文字で区切られている個々の文字列がそれぞれ番組検索のためのキーワードとして役立ち、タイトルそのものとしては限定されすぎて番組検索のためのキーワードとしてあまり役立たない（対戦チーム（□□や△△の具体名）が異なれば別のタイトルになってしまうので、検索結果がゼロまたは非常に少なくなる）ことが多い。したがって、ユーザーは、抽出されたキーワード（特殊文字で区切られている個々の文字列）を用いて、やはり番組を効率的に検索することができるようになる。

【0078】

他方、図4の処理では、フラッシュメモリ22に記憶されたEPG情報中のこれらのタイトルの番組の詳細文字列情報から、詳細情報用キーワード辞書に登録されている有名人（愛のから騒ぎという番組の司会者、ゲストや、映画スペース・ウォーズに出演している俳優）の名称等が、ステップS12でキーワードとして抽出される。

【0079】

その際、苗字が漢字で名前が平仮名やカタカナの有名人の名称（例えば石田あかりという名称）もこの詳細情報用キーワード辞書に登録されているので、そうした有名人の名称もキーワードとして抽出される。

【0080】

また、インターネット経由でダウンロードした最新の詳細情報用キーワード辞書も用いるので、最近有名になったばかりの人の名称もキーワードとして抽出される。

【0081】

また、その詳細文字列情報のうち、その有名人の名称等の部分と、詳細情報用除外文字列辞書に登録されている文字列（‘ゲスト’，‘以上’，‘監督’等）の部分とが、ステップS12及びS13で半角スペースに置換される。

【0082】

そして、このスペース置換された詳細文字列情報の文字列から、ステップS1

4で、図6に示したルールによってキーワードが抽出される。

【0083】

その際、カタカナとアルファベットとは同一の字種として扱われるとともに ‘.’ (中点) はその直前の文字がカタカナ、アルファベットである場合にはそれぞれカタカナ、アルファベットとして扱われるので、名前と苗字との間に ‘.’ (中点) が挿入された外国人名 (例えばB・ドゥーリー) もキーワードとして抽出される。

【0084】

また、最新の詳細情報用キーワード辞書にもまだ登録されていない人 (例えばデビューしたばかりの無名の芸能人) の名称でも、平仮名のみの名称や2文字以下の漢字のみの名称や6文字以上の漢字のみの名称 (すなわち人名としてあまりなさそうな名称) でなければキーワードとして抽出される。

【0085】

また、‘ゲスト’ , ‘以上’ , ‘監督’ といったような番組検索のためには不適切な文字列は、スペース置換されているのでキーワードとして抽出されることはない。

【0086】

これにより、ステップS15では、苗字が漢字で名前が平仮名やカタカナの有名な人名や、最近有名になったばかりの人の名称や、名前がアルファベットで表記され苗字がカタカナで表記された外国人名や、名前と苗字との間に ‘.’ (中点) が挿入された外国人名も、番組検索用のキーワードとしてフラッシュメモリ22に記憶される。したがって、ユーザーは、抽出されたキーワードを用いて、番組を効率的に検索することができるようになる。

【0087】

なお、図3、図4の処理によってフラッシュメモリ22に記憶させたキーワードをユーザーが番組検索のために用いる方法としては、例えば、リモコン28による所定の操作に基づき、CPU19が、番組検索用画面 (キーワードを一覧表示するとともにユーザーがその中の所望のキーワードを選択して検索を指示するための画面) の映像信号を作成して映像信号処理回路17、映像出力端子29を

経て表示装置 3 に送るといったような、適宜の方法をとればよい。

【 0 0 8 8 】

以上のように、この番組記録再生装置 2 では、E P G 情報中のタイトル文字列情報からのキーワードの抽出と詳細文字列情報からのキーワードの抽出とを、それぞれの情報に合せて互いに異なるキーワード辞書とルールとで行うことにより、小さなサイズのプログラムや辞書で精度よくキーワードを抽出することができるようになっている。

【 0 0 8 9 】

これにより、C P U 1 9 の処理能力やメモリ（R O M 2 0 やフラッシュメモリ 2 2 等）の容量がさほど大きくなくても、E P G 情報中のタイトル文字列情報及び詳細文字列情報から、ユーザーが番組を検索するためのキーワードを効率よく且つ精度よく自動的に抽出することができるようになっている。

【 0 0 9 0 】

なお、以上の例では、デジタルテレビジョン放送の番組を記録・再生する装置に本発明を適用している。しかし、これに限らず、アナログテレビジョン放送の番組を記録・再生する番組記録再生装置にも本発明を適用してよいことはもちろんである。

【 0 0 9 1 】

図 7 は、本発明を適用したアナログテレビジョン放送用の番組記録再生装置のハードウェア構成を示すブロック図である。アンテナ 3 1 で受信されて番組記録再生装置 4 1 に入力したアナログ放送信号中の映像・音声信号は、チューナ 4 2 で周波数帯を選択され、M P E G エンコーダ 4 3 で符号化される。

【 0 0 9 2 】

テレビ番組の視聴時には、この符号化された映像・音声データは、M P E G デコーダ 4 7 で復号されて、番組記録再生装置 4 1 から表示装置 6 1 に送られる。

【 0 0 9 3 】

他方、テレビ番組の記録時には、M P E G エンコーダ 4 3 で符号化された映像・音声データは、バス 4 4 を介して主記憶装置 4 5 に送られて、主記憶装置 4 5 に記録される。

## 【0094】

そして、再生時には、主記憶装置45から読み出された映像・音声データが、バス44を介してMPEGデコーダ47に送られ、MPEGデコーダ47で復号されて、番組記録再生装置41から表示装置61に送られる。

## 【0095】

また、チューナ42で周波数帯を選択されたアナログ放送信号から、EPG取得モジュール46でEPG情報が取得される。このEPG情報も、バス44を介して主記憶装置45に送られて、主記憶装置45に記憶される。

## 【0096】

また、インターネット71と接続するための通信インターフェース48、ROM49、主記憶装置50、補助記憶装置51、MPEGデコーダ47が、互いにバス52で結ばれている。

## 【0097】

この番組記録再生装置41でも、前述のようなタイトル用キーワード辞書、タイトル用除外文字列辞書、詳細情報用キーワード辞書及び詳細情報用除外文字列辞書がROM49に格納されている（詳細情報用キーワード辞書に関してはインターネット経由で専用のサイトから最新のものをダウンロードして補助記憶装置51にも記憶させる）とともに、番組記録再生装置41全体を制御するCPU53が、図3、図4に示したのと同じキーワードの自動抽出処理をこれらの辞書及び主記憶装置45内のEPG情報を用いて行い、抽出したキーワードを補助記憶装置51に記憶させる。

## 【0098】

この番組記録再生装置41でも、図1、図2の番組記録再生装置2について説明したのと全く同様にして、EPG情報中のタイトル文字列情報からのキーワードの抽出と詳細文字列情報からのキーワードの抽出とを、それぞれの情報に合せて互いに異なるキーワード辞書とルールとで行うことにより、小さなサイズのプログラムや辞書で精度よくキーワードを抽出することができる。

## 【0099】

これにより、CPU53の処理能力やメモリ（ROM49や補助記憶装置51

等)の容量がさほど大きくなくても、E P G情報中のタイトル文字列情報及び詳細文字列情報から、ユーザーが番組を検索するためのキーワードを効率よく且つ精度よく自動的に抽出することができる。

#### 【0100】

また、以上の例では、表示装置とは別体となった番組記録再生装置に本発明を適用している。しかし、これに限らず、この番組記録再生装置と表示装置とが一体となったテレビジョン受信機や、番組の記録再生機能を有しないテレビジョン受信機にも本発明を適用してよい。

#### 【0101】

また、以上の例では、E P G情報中の番組のタイトル文字列情報、詳細文字列情報からのキーワードの検索のために本発明を適用している。しかし、これに限らず、テレビジョン番組以外のコンテンツ（例えばインターネット経由で配信されるコンテンツ）のタイトル文字列情報、詳細文字列情報からのキーワードの検索のためにも本発明を適用してよい。

#### 【0102】

また、本発明は、以上の例に限らず、本発明の要旨を逸脱することなく、その他様々の構成をとりうることはもちろんである。

#### 【0103】

##### 【発明の効果】

以上のように、本発明によれば、C P Uの処理能力やメモリの容量がさほど大きくない家電製品でも、E P G情報のような番組のタイトル文字列情報及び詳細文字列情報から、ユーザーが番組を検索するためのキーワードを、効率よく且つ精度よく自動的に抽出することができるという効果が得られる。

##### 【図面の簡単な説明】

##### 【図1】

本発明を適用した番組記録再生装置を含むデジタルテレビジョン放送受信システムの概要を示す図である。

##### 【図2】

図1の番組記録再生装置のハードウェア構成を示すブロック図である。



【図 3】

図 2 の CPU が実行するキーワードの自動抽出処理を示すフローチャートである。

【図 4】

図 2 の CPU が実行するキーワードの自動抽出処理を示すフローチャートである。

【図 5】

図 3 の処理におけるキーワード抽出のためのルールを示す図である。

【図 6】

図 4 の処理におけるキーワード抽出のためのルールを示す図である。

【図 7】

本発明を適用したアナログテレビジョン放送用の番組記録再生装置のハードウェア構成を示すブロック図である。

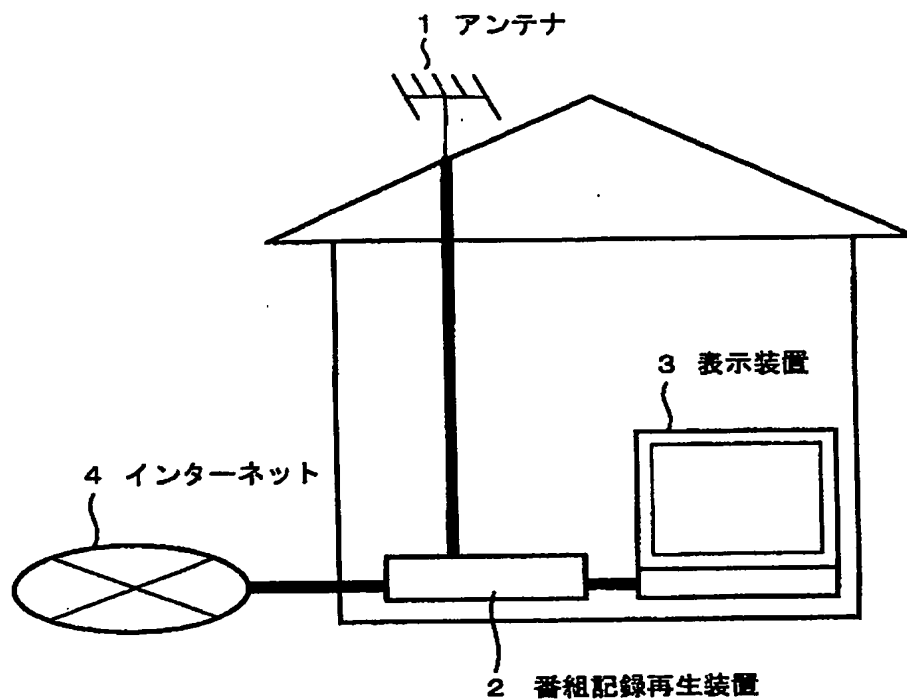
【符号の説明】

1, 31 アンテナ、 2, 41 番組記録再生装置、 3, 61 表示装置、 11, 42 チューナ、 12 復調器、 13 デスクランブラ、 14 多重分離器、 15 映像デコーダ、 16 音声デコーダ、 17 映像信号処理回路、 18 D/A変換器 18、 19 CPU、 20, 49 ROM、 21 RAM、 22 フラッシュメモリ、 23 リモートコントローラ用のインターフェース、 24 HDD用のインターフェース、 25, 48 インターネット接続用の通信インターフェース、 26 システムバス、 27 HDD、 28 リモコン、 43 MPEGエンコーダ、 44, 52 バス、 45, 50 主記憶装置、 46 EPG取得モジュール、 47 MPEGデコーダ、 50 主記憶装置、 51 補助記憶装置

【書類名】

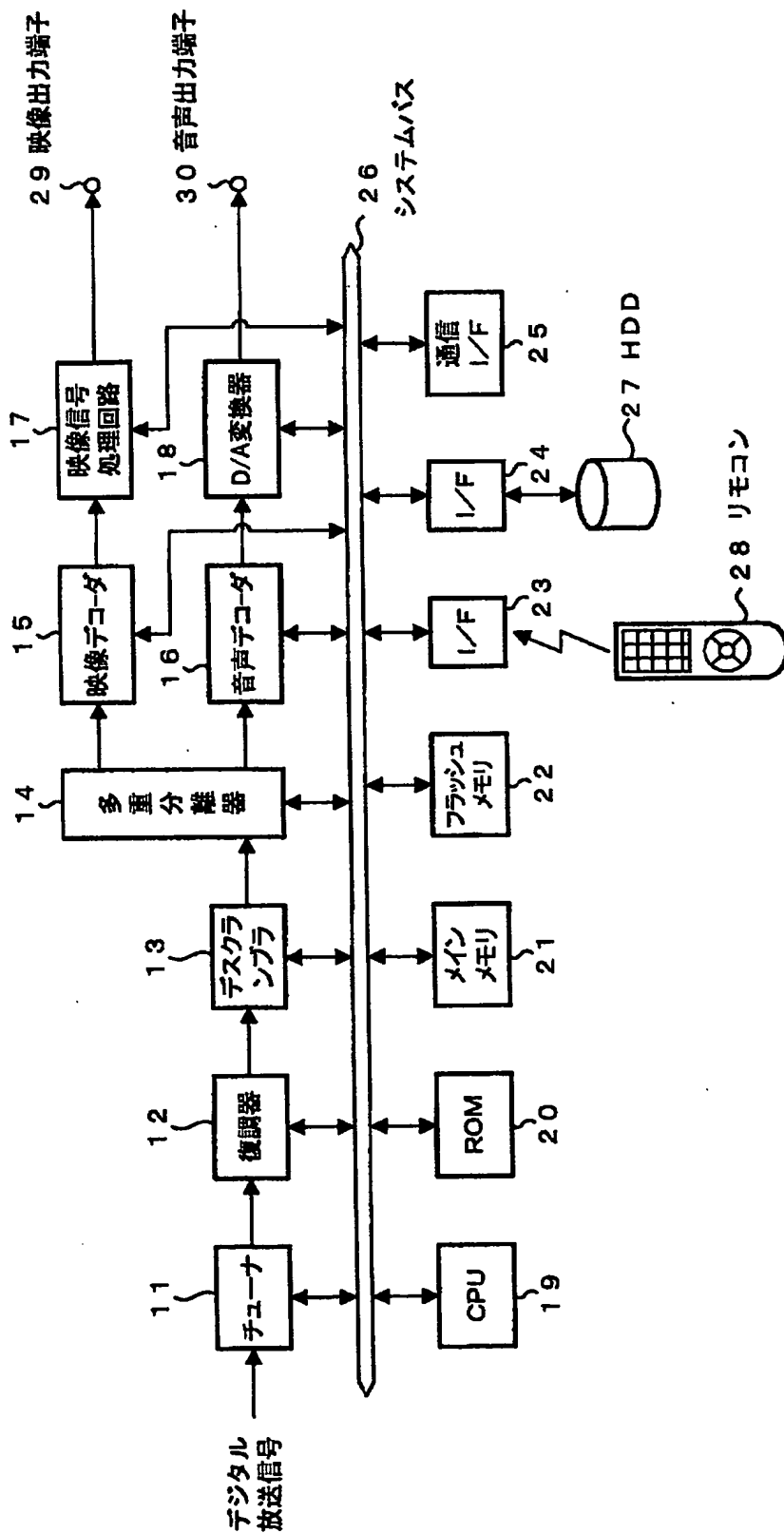
図面

【図 1】



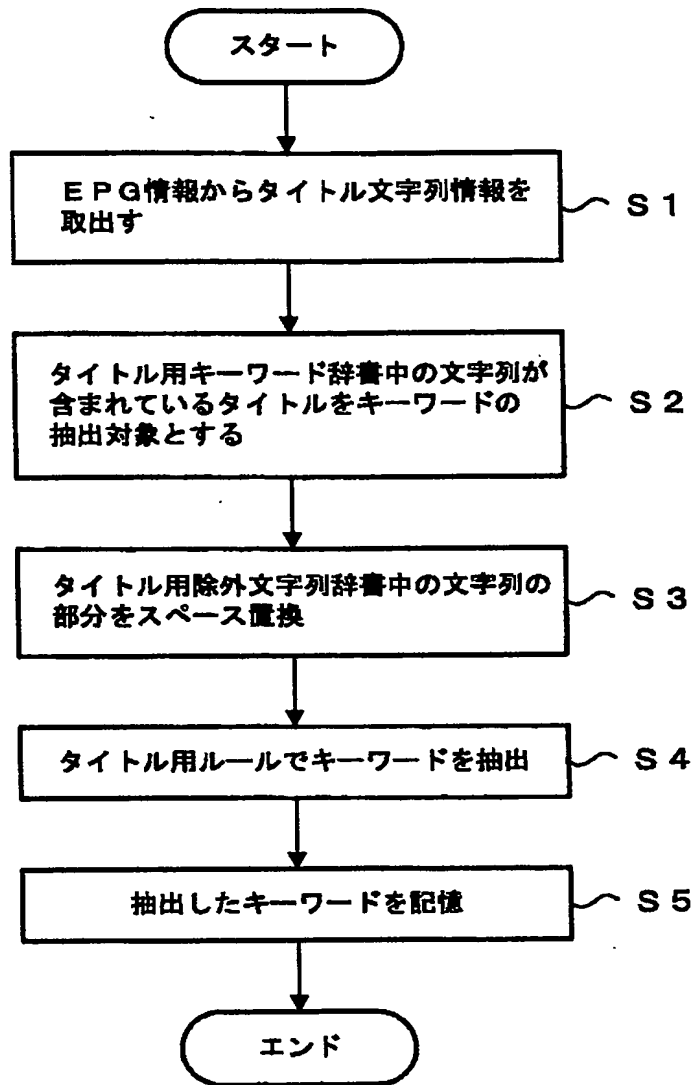
本発明の番組記録再生装置を含む放送受信システム

【図 2】



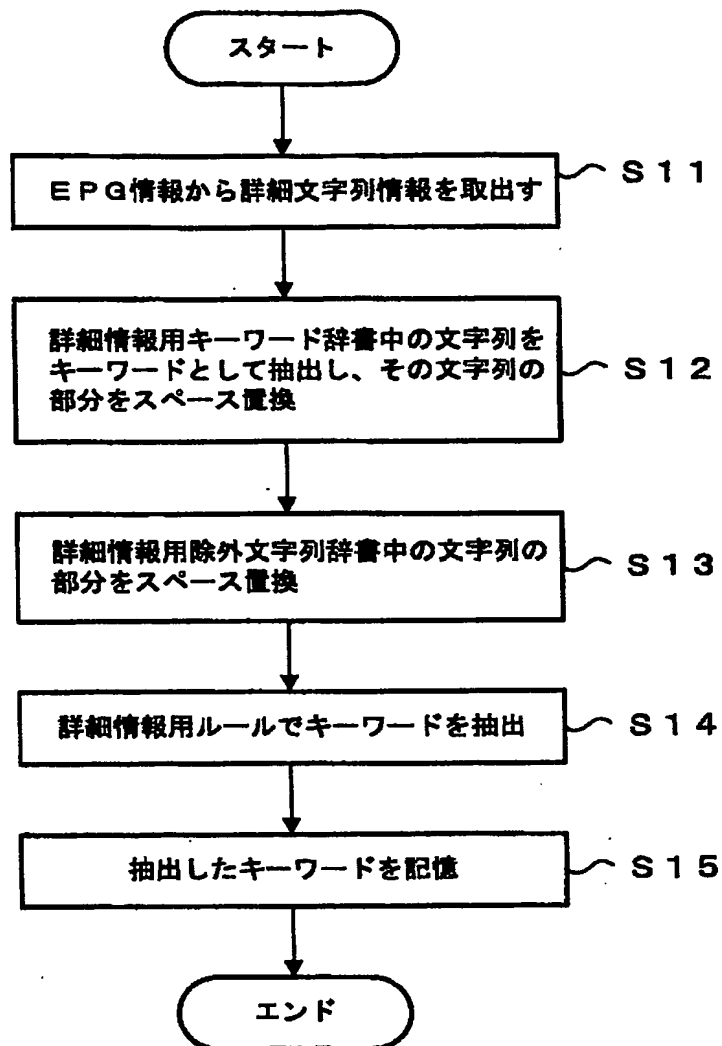
番組記録再生装置の構成

【図3】



キーワード自動抽出処理（タイトル文字列情報から）

【図4】



キーワード自動抽出処理（詳細文字列情報から）

【図 5】

特殊文字で区切られていないタイトルの 文字列をそのままキーワードとして抽出
特殊文字で区切られているタイトルの 各文字列のうち、2文字以上の文字列を キーワードとして抽出
‘ ’（中点）は特殊文字として扱わない。 抽出した文字列の先頭か末尾に、‘ ’が 存在する場合には、‘ ’を除いた部分を キーワードとする

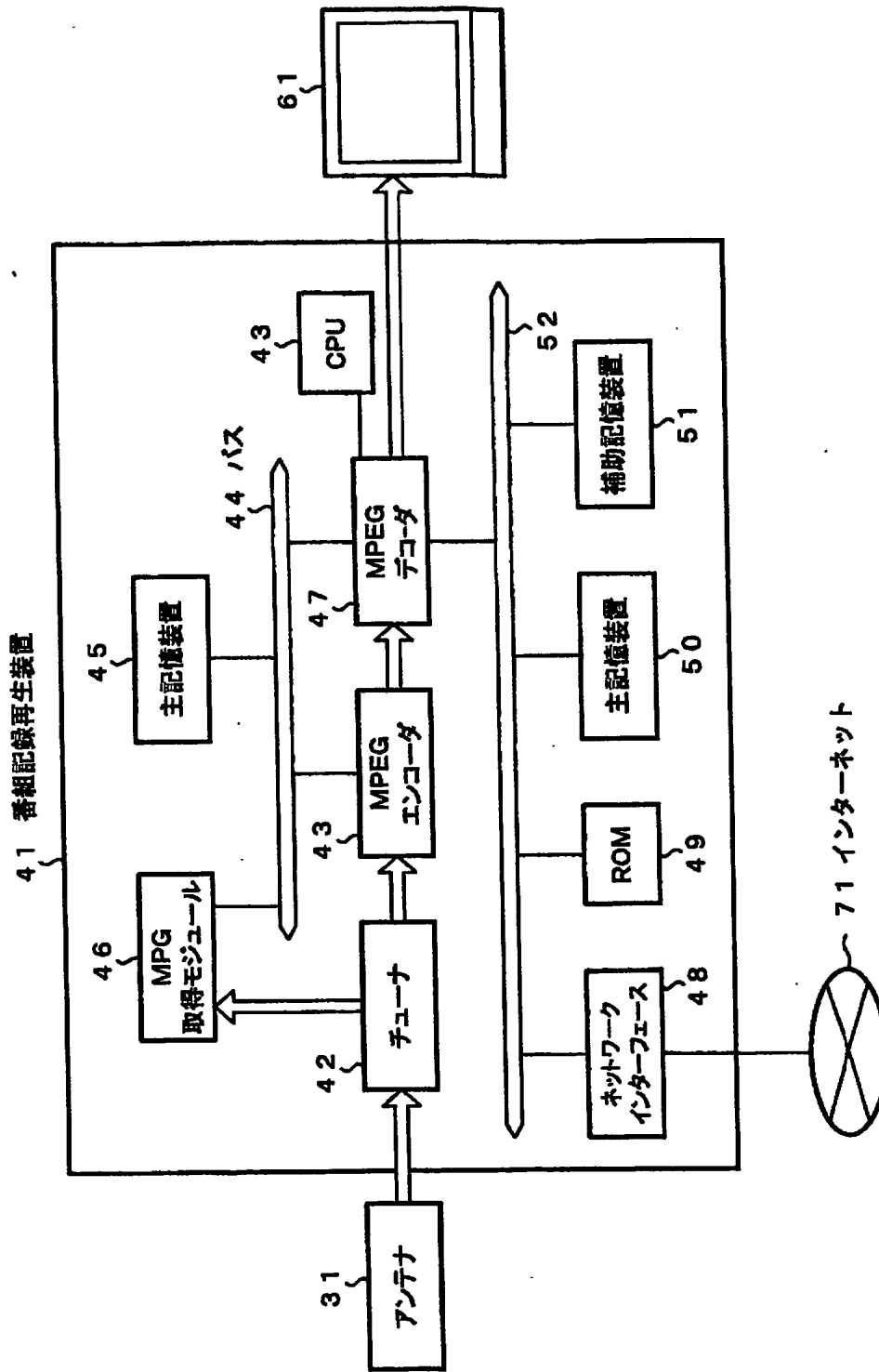
### キーワード抽出ルール（タイトル用）

【図 6】

字種切り法を利用
カタカナとアルファベットは同一字種として扱う
‘ ’は、直前の文字がカタカナ、アルファベット のとき、それぞれカタカナ、アルファベットとして 扱う
分離した文字列のうち、平仮名のみ、2文字以下の 漢字のみ、6文字以上の漢字のみを除いた文字列を キーワードとして抽出

### キーワード抽出ルール（詳細文字列情報用）

【図 7】



【書類名】 要約書

【要約】

【課題】 C P U の処理能力やメモリの容量がさほど大きくない家電製品でも、E P G 情報のようなコンテンツのタイトル情報及び詳細情報から、ユーザーがコンテンツを検索するためのキーワードを、効率よく且つ精度よく自動的に抽出できるようにする。

【解決手段】 コンテンツのタイトル情報から、サブジャンルを示す文字列が登録されている第 1 のキーワード辞書を用いてキーワードの抽出を行う第 1 の抽出手段 1 9 と、このコンテンツの詳細情報から、人名が登録されている第 2 のキーワード辞書を用いたキーワードの抽出と、字種切り法を利用したキーワードの抽出とを行う第 2 の抽出手段 1 9 とを備える。

【選択図】 図 2



認定・付加情報

特許出願の番号	特願 2002-221698
受付番号	50201125612
書類名	特許願
担当官	第一担当上席 0090
作成日	平成14年 7月31日

<認定情報・付加情報>

【特許出願人】

【識別番号】 000002185

【住所又は居所】 東京都品川区北品川6丁目7番35号

【氏名又は名称】 ソニー株式会社

【代理人】 申請人

【識別番号】 100122884

【住所又は居所】 東京都新宿区西新宿1丁目8番1号 新宿ビル  
信友国際特許事務所

【氏名又は名称】 角田 芳末

【選任した代理人】

【識別番号】 100113516

【住所又は居所】 東京都新宿区西新宿1丁目8番1号 新宿ビル  
松隈特許事務所

【氏名又は名称】 磯山 弘信

出 願 人 履 歴 情 報

識別番号 [000002185]

1. 変更年月日 1990年 8月30日  
[変更理由] 新規登録  
住 所 東京都品川区北品川6丁目7番35号  
氏 名 ソニー株式会社
2. 変更年月日 2003年 5月15日  
[変更理由] 名称変更  
住 所 東京都品川区北品川6丁目7番35号  
氏 名 ソニー株式会社